Final Exam - Moed B
Total time for the exam: 3 hours!

Please copy the following sentence and sign it: " I am respecting the rules of the exam: Signature:_________ "

Important: For **True / False** questions, copy the statement to your notebook and write clearly true or false. You should prove the statement if true, or disprove it, e.g. by providing a counter-example, otherwise.

1) **Transfer Entropy (36 Points):** Define the Transfer Entropy

$$\mathsf{TE}^{(k)}_{\mathcal{X}\to\mathcal{Y}}(t) = I\left(Y_t; X^{(k)}_{t-1} \middle| Y^{(k)}_{t-1}\right), \tag{1}$$

where $X^{(k)}_t := (X_t, X_{t-1}, ..., X_{t-k+1})$ is a notation for length-k history of a variable X up to time t.
Let $\{X_t\}$ and $\{Y_t\}$ be stationary and first-order Markov processes taking values from the binary alphabet:

- Process $\{X_t\}$ has a deterministic transitions from 0 to 1 or 1 to 0 each time step, i.e.

$$P(X_t|Y^{t-1}, X^{t-1}) = P(X_t|X_{t-1}), \quad P(X_t = x|X_{t-1} = x \oplus 1) = 1, \tag{2}$$

  where $P(X_0) \sim$ Bern $\left(\frac{1}{2}\right)$.
- Process $\{Y_t\}$ is a noisy observation of the last time step of $\{X_t\}$. Assume $\alpha \neq \frac{1}{2}$ and $0 < \alpha < 1$,

$$P(Y_t|Y^{t-1}, X^{t-1}) = P(Y_t|X_{t-1}), \quad P(Y_t = y|X_{t-1} = x) = \begin{cases} 1-\alpha & \text{if } y = x \\ \alpha & \text{if } y \neq x \end{cases}. \tag{3}$$

Reminder: A stochastic process $\{X_t\}$ is said to be **stationary** if for every t_1, t_2 and h, the joint probability distribution function $P(X_{t_1}, X_{t_1+1}, ..., X_{t_1+h})$ is equal to $P(X_{t_2}, X_{t_2+1}, ..., X_{t_2+h})$, i.e., the joint probability distribution is invariant under time shifts.

a) **(8 points) True / False** The described joint process $\{X_t, Y_t\}$ is stationary. Explain your answer.
b) **(6 points) True / False** $P(Y_t = y, X_{t-1} = x) \neq P(X_t = x, Y_{t-1} = y)$.
c) **(6 points)** Calculate the Mutual Information between Y_t and X_{t-1}, i.e. $I(Y_t; X_{t-1})$.
   Hint: Consider to use the fact that $Y_t = X_{t-1} \oplus Z_{t-1}$, where $\{Z_t\}$ are i.i.d. Bern(α).
d) **(6 points) True / False** $I(Y_t; X_{t-1}) = I(X_t; Y_{t-1})$.
e) **(6 points)** Show that the Transfer Entropy for $X \to Y$ with lag $k = 1$ is non-zero, i.e., $\mathsf{TE}^{(1)}_{\mathcal{X}\to\mathcal{Y}}(t) = I\left(Y_t; X_{t-1} \middle| Y_{t-1}\right) > 0$.
   Hint: Utilize the relation $Y_t = X_{t-1} \oplus Z_{t-1}$, and the fact that if $Z_1 \sim$ Bern(α) and $Z_2 \sim$ Bern(β), then $Z_1 \oplus Z_2 \sim$ Bern($\alpha - 2\alpha\beta + \beta$).
f) **(4 points)** Calculate the Transfer Entropy for $Y \to X$ with lag $k = 1$, i.e., $\mathsf{TE}^{(1)}_{\mathcal{Y}\to\mathcal{X}} = I\left(X_t; Y_{t-1} \middle| X_{t-1}\right)$.

2) **ML algorithms (36 Points):** Figure 1 shows the end-to-end communication system considered in this question. This system takes as input a bit sequence denoted by $\boldsymbol{b}$, which is then mapped onto symbols, $\mathrm{s} \in \mathcal{S}$. The sequence of symbols is fed into a symbol modulator that maps each symbol into a constellation point $x \in \mathbb{C}$. Both the modulator and demodulator are implemented with neural networks, hence they are learnable with trainable parameters $\boldsymbol{\theta}_M$ and $\boldsymbol{\theta}_D$, respectively. The demodulator maps each received sample $y \in \mathbb{C}$ to a probability vector $\tilde{p}_{\boldsymbol{\theta}_D}(s|y)$ over the set of symbols $\mathcal{S}$, as illustrated in Fig.1. Finally, the sent bits are reconstructed from $\tilde{p}_{\boldsymbol{\theta}_D}(s|y)$ by the symbols-to-bits mapper. Let us denote by $p_{\boldsymbol{\theta}_M}(s|y)$ the distribution induced by the system up to the point of the output channel (without the demodulator), which depends on the modulator parameters $\boldsymbol{\theta}_M$. We would like to approximate the true posterior distribution $p_{\boldsymbol{\theta}_M}(s|y)$ with the mapping defined by the demodulator $\tilde{p}_{\boldsymbol{\theta}_D}(s|y)$.
Given that the demodulator performs a classification task, the categorical cross-entropy is used as a loss function for training:

$$\mathcal{L}^*(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D) \triangleq \mathbb{E}_{S,Y}\{-\log\left(\tilde{p}_{\boldsymbol{\theta}_D}(S|Y)\right)\} \tag{4}$$

We assume that each symbol $\mathrm{s} \in \mathcal{S}$ is uniquely mapped to a constellation point $x \in \mathbb{C}$ (1:1 mapping), allowing us to replace s with x in expressions.

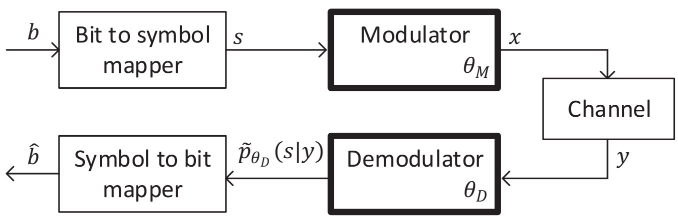

Fig. 1: Trainable end-to-end communication system. Trainable components are highlighted.

a) **(5 points)** In our system, $p_{\boldsymbol{\theta}_M}(x)$ represents the true distribution of x, and note that it depends on the modulator parameters, $\boldsymbol{\theta}_M$. The entropy of X under the distribution parameterized by $\boldsymbol{\theta}_M$ is:

$$H_{\boldsymbol{\theta}_M}(X) = -\sum_x p_{\boldsymbol{\theta}_M}(x)\log\left(p_{\boldsymbol{\theta}_M}(x)\right) \tag{5}$$

True/False: $H_{\boldsymbol{\theta}_M}(X) = H(S)$? Explain why.

b) **(5 points)** Given a sequence of i.i.d. samples s_i and y_i over a long period of time, for $i = 1, 2, 3, ...$, how can you compute $\mathcal{L}^*(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D)$?

c) **(5 points)** Recall the 1:1 mapping between s to x. Let:

$$\mathcal{L}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D) \triangleq \mathbb{E}_Y\{H\left[p_{\boldsymbol{\theta}_M}(x|y), \tilde{p}_{\boldsymbol{\theta}_D}(x|y)\right]\} \tag{6}$$

True/False $\mathcal{L}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D)$ equals $\mathcal{L}^*(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D)$, where $\mathcal{L}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D)$ is defined in (6), and $\mathcal{L}^*(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D)$ is defined in (4). If yes, prove it. Otherwise, correct the equation, and explain your reasoning.

d) **(5 points) True/False** The loss function, $\mathcal{L}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D)$, as defined in (6), satisfies the following equality:

$$\mathcal{L}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D) = -\sum_x\sum_y p_{\boldsymbol{\theta}_M}(x)p_{\boldsymbol{\theta}_M}(y|x)\log\left(\tilde{p}_{\boldsymbol{\theta}_D}(x|y)\right). \tag{7}$$

If yes, explain why. If no, provide the correct expression.

e) **(6 points) True/False** The loss function can be expressed as the following equation:

$$\mathcal{L}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D) = H(S) - I_{\boldsymbol{\theta}_M}(X;Y) + \mathbb{E}_Y\{D_{KL}\left(p_{\boldsymbol{\theta}_M}(x|y)\|\tilde{p}_{\boldsymbol{\theta}_D}(x|y)\right)\}. \tag{8}$$

If yes, prove it. Otherwise, correct the equation and explain your reasoning.

f) **(5 points)** A student who saw equation (8) claims that:

$$\arg\min_{\boldsymbol{\theta}_M, \boldsymbol{\theta}_D} \hat{\mathcal{L}}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D) = \arg\min_{\boldsymbol{\theta}_M, \boldsymbol{\theta}_D} \mathcal{L}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D) \tag{9}$$

where $\hat{\mathcal{L}}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D)$ is defined as:

$$\hat{\mathcal{L}}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D) = \mathcal{L}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D) - H(S) \tag{10}$$

Is the student claim **True/False**? Please explain/justify your answer.

g) **(5 points)** Please explain the contribution/meaning of minimizing the first loss component of $\hat{\mathcal{L}}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D)$, namely $-I_{\boldsymbol{\theta}_M}(X;Y)$, and the second loss component of $\hat{\mathcal{L}}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D)$, namely $\mathbb{E}_y\{D_{KL}\left(p_{\boldsymbol{\theta}_M}(x|y)\|\tilde{p}_{\boldsymbol{\theta}_D}(x|y)\right)\}$, to the overall communication system.

3) **Polar compressor (32 Points):** For a positive integer N, let $n = 2^N$ and consider the invertible matrix $P_n \in \mathbb{F}_2^{n\times n}$ defined by:

$$P_n = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^{\otimes N}.$$

Further, consider $Z^n = (Z_1, \ldots, Z_n) \sim \text{Bern}(p)^n$ where $p \in (0, 0.5)$, and let $W^n = Z^n \cdot P_n$.

a) **(4 points)** For both channel coding and source coding, we do polarization. Explain briefly the difference in polarization between channel coding and source coding.

b) **(4 points)** Define the rate for source coding and channel coding, and explain whether you want to maximize or minimize those rates.

c) **(6 points)** Assume $n = 4$ and consider the entropy terms $H(W_i|W^{i-1})$ for $i \in \{1, \ldots, 4\}$. Determine and explain which one is the highest, and calculate this specific entropy term explicitly in terms of p.

d) **(6 points)** Define the set S_τ as follows:

$$S_\tau = \{i \in \{1, \ldots, 4\} \mid H(W_i|W^{i-1}) \geq \tau\}.$$

For $\hat{\tau} = -\mathbb{E}[\log_2(P_{W_1})]$, write explicitly the set $S_{\hat{\tau}}$. Explain your result.

e) **(6 points)** Consider $z^4 = [1, 0, 1, 1]$ and the set $S_{\hat{\tau}}$ that you found in the previous item. What is the output of the encoder?

f) **(6 points)** This time let $n = 2$, and assume that $Z_1 \sim \text{Bern}(p_1)$ and $Z_2 \sim \text{Bern}(p_2)$ are sampled conditioned on $Z_1 + Z_2 = a$ (for $a \in \mathbb{F}_2$). Let $b(p_1, p_2, a)$ denote the probability of Z_2 being 1 conditioned on $Z_1 + Z_2 = a$. Find $b(p_1, p_2, a)$ for both $a = 0$ and $a = 1$.

Good Luck!